**Author Manuscript**
Peer reviewed and accepted for publication by a journal

# A comparison of unamplified and massively multiplexed PCR amplification for murine antibody repertoire sequencing

Trisha A Rettig [*], Michael J Pecaut [†], Stephen K Chapes [*,1]

Author information   Article notes   Copyright and License information
PMCID: PMC6366624   NIHMSID: NIHMS987236   PMID: 30740592

## Abstract

Sequencing antibody repertoires has steadily become cheaper and easier. Sequencing methods usually rely on some form of amplification, often a massively multiplexed PCR prior to sequencing. To eliminate potential biases and create a data set that could be used for other studies, our lab compared unamplified sequencing results from the splenic heavy-chain repertoire in the mouse to those processed through two commercial applications. We also compared the use of mRNA vs total RNA, reverse transcriptase, and primer usage for cDNA synthesis and submission. The use of mRNA for cDNA synthesis resulted in higher read counts but reverse transcriptase and primer usage had no statistical effects on read count. Although most of the amplified data sets contained more antibody reads than the unamplified data set, we detected more unique V-gene segments in the unamplified data set. Although unique CDR3 detection was much lower in the unamplified data set, RNASeq detected 98% of the high frequency CDR3s. We have shown that unamplified profiling of the antibody repertoire is possible, detects more V-gene segments, and detects high frequency clones in the

repertoire.

**Keywords:** High throughput sequencing, Immunoglobulin genes, Mouse, Heavy Chain, RepSEQ

## Introduction

Antibodies are B cell proteins that play a vital role in adaptive immunity. These complex molecules, and their diverse specificities are defined at several levels. At the protein level, these molecules are heterodimers joined with disulfide bonds between heavy and light chains. At the genetic level, the antibody specificity is influenced by semi-randomly recombining a variable (V), diversity (D), and joining (J) gene segments that are encoded in the genome. This V-D-J, or in the case of the light chain, V-J rearrangement, also influences antibody specificity by random and semi-random base insertions or deletions during the recombination process. In the end, the collection of antibodies produced, the antibody repertoire, is a fingerprint of what antigens an organism has been exposed to and a measure of immunocompetence. Sequencing has become easier, cheaper, and faster in recent years. Antibody repertoires have been sequenced in numerous species, in response to vaccinations (1–3) and infections (4–6), and have been employed in cancer detection; providing valuable feedback regarding the immune system's response to challenges and for early cancer detection in patients (7, 8).

In an effort to supplement unamplified MiSeq data sets used in our laboratory (9), we explored the use of commercial processes that use technologies to amplify sequences to increase the depth of coverage of specifically targeted immunoglobulin gene transcripts. These data sets are created using massively multiplexed PCR reactions that are subsequently sequenced on the Illumina platform. Multiplex amplification strategies have been used to explore the T cell (10, 11) and B cell repertoires (12–15) and biases, omissions (10, 11, 16–18).and PCR artifacts have been detected in the data sets (10).

We started this project with the hypothesis that our unamplified data set would provide comparable results to those seen in the commercially amplified data sets. However, we are unaware of any immunoglobulin repertoire studies that have been done to compare data obtained using amplification techniques compared to the repertoire in a total RNASeq library. Concurrent with performing these analyses, we discovered that the required sample preparation for commercial sequencing also varied. Therefore, we found it necessary to examine the impact of sample preparation as part of our effort. This manuscript describes our results comparing a data set generated using unamplified total RNA (TRNA) to commercially amplified data sets. We examined the role of commercial amplification and cDNA generation methods as well as the impact of the starting material on sequence output.

## Materials and Methods

## KSU RNA Preparation

RNA was prepared as outlined in Rettig et al 2017 (9). Briefly, RNA was extracted from the spleens of four nine-week old female C57Bl/6J mice. TRNA was submitted to the Kansas State University Integrated Genomics Facility for sequencing and cDNA was prepared using standard Illumina protocols. cDNA made from the TRNA using random hexamer primers and oligo-dT selection, was size selected to 275-800bp length and sequenced on the Illumina MiSeq at 2×300bp using Illumina's protocol. No additional amplification beyond that required by Illumina preparation was used and consider these samples non-specifically amplified, or "unamplified".

## Commercial Sample preparations

mRNA was extracted from the TRNA isolated from the sample used for the KSU data set using the PolyATtract mRNA isolation system (Promega, Fitchburg, WI) following manufacturer's instructions. RT-PCR for samples amplified with Avian Myeloblastosis Virus (AMV) reverse transcriptase based sample preparation was performed using the Access RT-PCR System (Promega, Fitchburg, WI) following the manufacturer's instructions. RT-PCR for Moloney Murine Leukemia Virus (MMLV) reverse transcriptase based sample preparation was performed using the SuperScript III First-Strand Synthesis System (Invitrogen, Carlsbrad, CA) following manufacturer's instructions. Starting material for RT-PCR was either TRNA from the KSU data set or the purified mRNA. RT-PCR products were submitted to Adaptive Biotechnologies (Seattle, WA) **(Com1)** on dry ice following company protocols. TRNA and mRNA (unamplified) were submitted to iRepertoire, Inc. (Huntsville, AL) **(Com2)** on dry ice following company protocols.

## Bioinformatic Analysis

KSU sequencing results were analyzed as described previously (9). Briefly, sequencing results were quality controlled. Antibody-specific sequences were isolated and submitted to ImMunoGeneTics (IMGT) (19) for analysis. Individual sequences were assigned a unique ID by the sequencing machinery during Illumina sequencing and were used to identify unique sequences. The sequence containing the most high quality information was retained for further analysis as outlined in Rettig *et al* (9). No further filtering of reads was performed. Both commercial sequencers provided their own bioinformatic analyses of the sequencing results. The raw sequencing results from Com1 and Com2 were also submitted to IMGT for analysis and subjected to the standard KSU bioinformatics pipeline. IMGT's nomenclature and classifications were used throughout this paper. We assessed all functional V-gene segments as identified by IMGT. We also include three putative functional genes (V5S21, V1S100, and V3S7) which were detected in rearranged transcripts (containing a CDR3 C-xx-W motif or class switched) in our previous analysis of the normal C57BL/6 repertoire (20). IMGT's High-V Quest occasionally assigned multiple potential V-gene segments to a single sequence, likely due to incomplete capture of the entire V-gene sequence or high homology between gene segments.

In all IMGT processed data, sequences that contained two possible V-gene segment possibilities were assigned a weighted value of 0.5 per sequence, as opposed to one for full matches. Sequences with V-gene segments that were assigned more than two potential matches were excluded from analysis. Initial results were tabulated using the companys' proprietary bioinformatic results. However, to determine the role of bioinformatic handling of the data, some of Com1 and Com2 data were subjected to the standard KSU bioinformatic workflow analysis and CDR3 analyses (9).

## Statistical Tests

All statistical analyses were carried out using GraphPad Prism (Version 6.0). Paired T-tests were performed using the raw read counts. Coefficient of determinations ($R^2$) were performed by comparing the percent of repertoire between animals. Percent of repertoire is determined by dividing the read count for a specific V-gene segment by the total number of reads detected and multiplying by 100.

## Results

Most studies examining immunoglobulin repertoires use amplification to increase the depth of sequencing, but amplification comes with some drawbacks. We wanted to assess the comparability of amplified and non-amplified data from identical samples. In preparation to do this comparison, we found that different commercial amplification methodologies required different types of sample preparation. For example, sample submission for the Com1 data sets required a cDNA sample. The Com1 process amplified the resulting cDNA using proprietary primers and sequencing on the Illumina platform. After an initial submission showed a low correlation between the Com1 sequencing and the KSU data set (data not shown), we hypothesized that cDNA preparation plays a role in determining the amplified repertoire. To test this hypothesis, we assessed the role of starting material (mRNA or TRNA), reverse transcriptase (AMV vs MMLV), and primer templates (oligo-dT or random hexamer) on the sequenced B cell immunoglobulin repertoire. Com2 submissions required the submission of TRNA, rather than cDNA.

### Assessment of Transcriptional Read Counts

Com1amplified data sets returned between 7,084 and 1,263,003 sequences, dependent on the preparation method. mRNA starting material yielded more total transcriptional reads than TRNA (P=0.013, 2 tailed matched T test; Table 1). Generally, the AMV reverse transcriptase and random hexamer primers tended to yield higher numbers of transcripts. The use of AMV and random hexamer primers resulted in more total productive reads in three out of four of the comparisons directly comparing primers, however, the overall differences were not statistically different (P>0.05, 2 tailed matched T test; Table 1). In the Com2 data set, we found a moderate number of reads, about one-half of those detected in the highest Com1 numbers. These compare to 11,200 sequence reads containing a CDR3 generated in the KSU data set using a total MiSeq approach.

## Table 1:

Total number of productive reads per data set

| | KSU[a] | Com 1[a] | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | mRNA[b] | | | | TRNA[b] | | | |
| | | AMV[c] | | MMLV[c] | | AMV[c] | | MMLV[c] | |
| | | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] |
| Total Productive Reads | 11,200[e] | 553,521 | 1,263,003 | 883,532 | 1,035,461 | 7,975 | 6,867 | 208,979 | 220,772 |

Open in a new tab

[a] – Sequencing technique (Com1 and Com2 are amplified data sets)

[b] – starting material (mRNA – Messenger RNA, TRNA – Total RNA)

[c] – reverse transcriptase (AMV - Avian Myeloblastosis Virus, MMLV - Moloney Murine Leukemia Virus)

[d] – primer (dt – Oligo dT, Hex – Random hexamer)

[e] – an additional 27,896 reads were used for V-gene segment usage assessment. These sequences were not long enough for CDR3 detection.

## Determination of Sequencing Reproducibility

To assess the repeatability of the amplified Com1 and Com2 data sets, we examined the correlation of V-gene segment usage. In the C57BL/6 mouse, the V-gene segment is the most varied in the heavy chain (IgH locus) comprising a total of 109 functional V-gene segments, three putative function V-gene segments, and alleles compared to 19 for the D-gene segment and four for the J-gene ([21], [22]). The nucleotide sequences in the V-gene segments are also highly varied and require a complex multiplex PCR to amplify. Correlations were assessed using the data provided by the commercial vendor's proprietary bioinformatics.

Non-strain specific V-gene segment assignments accounted for between 0.84% and 1.46% of the sequencing results from Com 1 and 1.74% and 1.41% for Com2 (Table 2). Although there were differences in the immunoglobulin gene

transcripts detected, there was a high correlation in the V-gene sequences detected among the different technical replicates in the Com1 data ($R^2$ range from 0.6986 to 0.9933, all $p<0.0001$) ([Figure 1](#)). The $R^2$ between technical replicates in Com2 was 0.9621 ($p<0.0001$). We also examined the reproducibility of two technical replicates of KSU sequencing used for a different analysis and obtained an $R^2$ of 0.9996 for Mouse 32, and 0.9995 for Mouse 39 ($p<0.001$) showing high levels of reproducibility between KSU sequencing runs. Therefore, although total transcripts generated varied with sample preparation, the V-gene segments that were detected were consistently detected using two different commercial approaches.

## Table 2:

Percent of non-C57BL/6 V-gene segments detected per data set

| | Com 1[a] | | | | | | | | Com |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | mRNA[b] | | | | TRNA[b] | | | | mRNA[b] |
| | AMV[c] | | MMLV[c] | | AMV[c] | | MMLV[c] | | |
| | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] | |
| % Non-B6 V-Gene segments | 0.92 | 0.88 | 0.84 | 1.17 | 1.30 | 1.46 | 1.02 | 1.14 | 1.74 |

[Open in a new tab](#)

a – Sequencing technique (Com1 and Com2 are amplified data sets)
b – starting material (mRNA – Messenger RNA, TRNA – Total RNA)
c – reverse transcriptase (AMV - Avian Myeloblastosis Virus, MMLV - Moloney Murine Leukemia Virus)
d – primer (dt – Oligo dT, Hex – Random hexamer)

# Figure 1.

| | | Com 1 | | | | | | | Com2 |
|---|---|---|---|---|---|---|---|---|---|
| | | mRNA AMV Hex | mRNA MMLV dT | mRNA MMLV Hex | Total AMV dT | Total AMV Hex | Total MMLV dT | Total MMLV Hex | mRNA |
| **Com 1** | mRNA AMV dT | 0.9933 | 0.9185 | 0.9547 | 0.8867 | 0.9342 | 0.9607 | 0.9006 | |
| | mRNA AMV Hex | | 0.9119 | 0.9575 | 0.8888 | 0.9392 | 0.9607 | 0.9068 | |
| | mRNA MMLV dT | | | 0.9335 | 0.8526 | 0.8090 | 0.9261 | 0.8076 | |
| | mRNA MMLV Hex | | | | 0.8016 | 0.8803 | 0.9650 | 0.9395 | |
| | Total AMV dT | | | | | 0.8558 | 0.8264 | 0.6986 | |
| | Total AMV Hex | | | | | | 0.9040 | 0.8640 | |
| | Total MMLV dT | | | | | | | 0.9262 | |
| **Com2** | TRNA | | | | | | | | 0.9621 |

Open in a new tab

R2 values of sequencing technical replicates. The percent of repertoire for each V-gene segment detected was compared between technical replicates. The highest R2 are dark read, the lowest R2 are blue.

## Impact of Amplification on V-gene Segment Detection

The unamplified KSU approach produced a data set where a total of 112 V-gene segments were detected while the Com1 data sets contained between 85 and 100 V-gene segments. The Com2 mRNA data set contained 99 detectable V-gene segments and the TRNA contained 100. Comparisons of V-gene segments in the Com1 data set to the KSU data set showed moderate $R^2$ values; 0.4457 to 0.5841 (all p<0.0001) ([Table 3](#)). The Com2 data sets also showed moderate $R^2$ of 0.6695 for mRNA and 0.6607 for TRNA (all p<0.0001). To determine why there were differences in V-gene detection, we compared the results from the most commonly detected V-gene segments in the KSU data sets to their frequencies in the Com1 and Com2 data sets. The protocol for Illumina sequencing uses mRNA selection, SuperScriptIII reverse transcriptase, and random hexamer primers. To best compare results, we used the Com1 mRNA-MMLV-Hex data set and the Com2 mRNA data set using the top 34 V-gene segments in the KSU data set. These V-gene segments comprise over one percent of the detected repertoire and are considered highly expressed. In the Com1 data set, of these 36 highly expressed V-gene segments, five (V1-26, V1-18, V1-50, V4-1, and V2-6) were detected at two-fold lower frequency than in the KSU data set ([Figure 2](#)). These five V-gene segments were either absent or were near zero percent of the repertoire ([Figure 2](#)). Of these same top 34 V-gene segments in the Com1 data set, two (V6-3 and V2-6-8) were detected at two-fold greater than the KSU data set ([Figure 2](#)). The $R^2$ for these top 36 V-gene segments to the KSU data set was 0.1783 (P=0.0128).

## Table 3:

Correlations of data sets to unamplified KSU data set and read counts

| | Com 1[a] | | | | | | | | Com |
| | mRNA[b] | | | | TRNA[b] | | | | mRNA[b] |
| | AMV[c] | | MMLV[c] | | AMV[c] | | MMLV[c] | | |
| | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] | dT[d] | Hex[d] | |
| $R^2$ to KSU Dataset | 0.5677 | 0.5773 | 0.4496 | 0.5517 | 0.4457 | 0.5606 | 0.5554 | 0.5841 | 0.6695 |
| Assessed V-Gene Segments | 506,503 | 151,104 | 1,749,618 | 1,245,999 | 267,946 | 5,666 | 267,946 | 302,057 | 626,093 |

Open in a new tab

[a] – Sequencing technique (Com1 and Com2 are amplified data sets)

[b] – starting material (mRNA – Messenger RNA, TRNA – Total RNA)

[c] – reverse transcriptase (AMV - Avian Myeloblastosis Virus, MMLV - Moloney Murine Leukemia Virus)

[d] – primer (dt – Oligo dT, Hex – Random hexamer)

Figure 2.

| | KSU | Com1 | Com2 | Prod |
|---|---|---|---|---|
| V1-80 | 7.10 | 6.07 | 6.40 | 7.10 |
| V1-26 | 4.88 | 0.00 | 6.31 | 3.70 |
| V1-9 | 3.59 | 4.81 | 5.04 | 3.71 |
| V1-18 | 3.14 | | 2.17 | 2.81 |
| V1-53 | 3.05 | 3.95 | 6.55 | 3.06 |
| V6-3 | 2.78 | 5.67 | 1.77 | 2.53 |
| V9-3 | 2.76 | 2.76 | 0.92 | 2.82 |
| V1-55 | 2.44 | 1.97 | 4.26 | 2.02 |
| V1-50 | 2.37 | | 2.00 | 2.01 |
| V3-6 | 2.31 | 3.09 | 4.99 | 1.93 |
| V1-76 | 2.26 | 3.11 | 4.08 | 2.73 |
| V6-6 | 2.26 | 3.31 | 1.74 | 2.20 |
| V1-64 | 2.12 | 2.82 | 4.64 | 2.29 |
| V4-1 | 2.08 | 0.00 | 0.01 | 1.93 |
| V2-6 | 2.01 | 0.01 | 0.01 | 2.20 |
| V2-6-8 | 2.01 | 4.49 | 1.93 | 2.20 |
| V1-82 | 1.99 | 1.95 | 1.72 | 2.51 |
| V1-22 | 1.81 | 1.88 | 2.42 | 1.76 |
| V1-81 | 1.79 | 2.07 | 0.04 | 1.73 |
| V5-17 | 1.78 | 3.01 | 0.61 | 1.81 |
| V2-2 | 1.77 | 1.99 | 0.67 | 2.22 |
| V8-8 | 1.77 | 2.47 | 0.67 | 2.38 |
| V7-3 | 1.66 | 1.06 | 2.66 | 1.56 |
| V10-1 | 1.64 | 2.20 | 4.32 | 1.28 |
| V1-72 | 1.56 | 2.24 | 2.35 | 1.25 |

| | | | | |
|---|---|---|---|---|
| V10-3 | 1.49 | 1.51 | 1.75 | 1.21 |
| V1-69 | 1.43 | 0.97 | 3.45 | 1.42 |
| V2-3 | 1.42 | 1.96 | 2.39 | 1.52 |
| V11-2 | 1.41 | 2.43 | 0.12 | 1.97 |
| V14-4 | 1.32 | 1.16 | 1.79 | 1.96 |
| V1-78 | 1.19 | 1.04 | 1.18 | 1.47 |
| V1-19 | 1.10 | 1.57 | 1.65 | 1.19 |
| V8-12 | 1.04 | 1.18 | 0.54 | 1.37 |
| V14-2 | 1.01 | 1.33 | 0.11 | 1.41 |

Percent of repertoire for high frequency V-gene segments among data sets. Percent of repertoire for the KSU, Com1 (mRNA-MMLV-hex), Com2 (mRNA), and Prod (productive only sequences from the KSU data set) are displayed. The highest value percent of repertoire is dark read while the lowest are white. Black boxes represent no detected reads (true zero). Rounded zeros are represented as 0.0.

We also compared the Com2 mRNA data set to the same 34 V-gene segments from the unamplified KSU data. Nine of the 34 V-gene sequences (V9-3, V4-1, V2-6, V1-81, V5-17, V2-2, V8-8, V11-2, V14-2) were detected at a twofold or lower level than in the KSU data set (Figure 2). Five other V-gene segments (V1-53, V3-6, V1-64, V10-1, and V1-69) were detected at twofold higher levels than those found in the KSU data set (Figure 2). The correlation of these top 34 V-gene segments was better than the Com1 $R^2$ at 0.4098 (P=<0.0001).

Our normal workflow methods include the use of functionally productive and unknown transcripts for analysis (9). This inclusion helps balance the lower read numbers obtained with unamplified sequences. We performed the same analysis as above between our productive + unknown dataset used above, with our productive only dataset. We detected a total of 104 V-gene segments. Those not detected in the productive only list (V3S7, V6-7, V6-4, V1-62-1, V5-12-4, V1-17-1, and V6-5) comprised less than 0.7% of the repertoire. The correlation coefficient was high at 0.9596(P<0.0001), and there were no changes at great than 2-fold of the productive + unknown data set (Figure 2). These analyses reveal that the addition of unknown functionality V-gene segments does not significantly alter the repertoire.

## Direct Comparisons of Amplified and Unamplified Data Sets

The comparisons in V-gene use were made using the bioinformatics provided by the commercial ventures. To standardize the data handling to remove bioinformatic reasons for the differences in data, we processed the sequencing results from the Com1 mRNA-MMLV-Hex and Com2 mRNA data sets using the KSU bioinformatics work flow (9).

The KSU bioinformatic treatment of the Com1 data set correlated moderately with the commercially provided bioinformatics ($R^2$=0.4795, p<0.0001). After processing the Com1 data with the KSU bioinformatics pipeline, the $R^2$ to the KSU data set increased slightly from 0.5517 (Table 3) with the original bioinformatics to 0.5649 (p<0.0001) with the adjusted bioinformatics. However, nine V-gene segments were detected in the Com1 data set using the KSU bioinformatics workflow that were not originally detected using the commercially provided bioinformatics (Sup Fig 1). When we processed the Com2 data using the KSU bioinformatic pipeline, the Com2 data set was highly correlated with the original commercially provided bioinformatics treatment ($R^2$=0.9860, p<0.0001). When we compared Com2 data set processed with the KSU bioinformatics pipeline to the KSU RNASeq data set, the data still only had an $R^2$=0.6791 (p<0.0001). The KSU bioinformatics workflow detected an additional four V-gene segments that were not detected by the commercial bioinformatics (Sup Fig 1).

When we reanalyzed the bioinformatics data from Com1 and Com2 using the KSU pipeline, we detected gene segments that were not detected in the original commercially provided bioinformatics. However, the inclusion of these gene segments, did not greatly improve the $R^2$ between the amplified data sets and the KSU RNASeq data. In the Com1 data set, some gene segments (V1-26, V1-18, V1-50, V2-9-1) were not detected or only detected at low levels in the original bioinformatics but were detected at high levels (>1%) in the KSU/IMGT processed data (Sup Fig 1). The three additional V-gene segments detected in the Com2 data set (V2-5, V1-62-2, and V1-62-3 were found in less than <0.3% of the repertoire (Sup Fig 1). These changes were not sufficient to significantly improve $R^2$ values.

## Impact of Amplification on the Reproducibility of CDR3 Detection

The absence of some V-gene segments in the Com1 and Com2 data compared to the KSU data was a concern. It precludes a complete picture of the V-gene repertoire. Nevertheless, amplified sequencing of the antibody repertoire is thought to provide an advantage in that the depth of coverage is increased over unamplified data sets due to the number of reads generated. To determine how extensive the discrepancy is between amplified and unamplified data, we assessed the read depth (number of reads generated) and resampling efficiency of CDR3 (number of unique CDR3s resampled between replicates) using technical replicates of samples sequenced with the various sequencing techniques. As anticipated, amplified data sets had both higher total read numbers and unique CDR3 numbers (Table 4).

Table 4:

| | KSU | mRNA-MMLV-Hex (Com1) | mRNA (Com2) |
|---|---|---|---|
| Read Count | 11,200 | 1,035,461 | 637,214 |
| Unique CDR3 Sequences | 6668 | 180,266 | 146,231 |

Resampling/reproducibility has been assumed to improve with the depth of coverage. We had the unique opportunity to compare sequencing data from the same biological material subjected to multiple sequencing methodologies. We have also had the opportunity to do technical replicates on multiple samples subjected to RNASeq or amplification procedures. This allowed us the ability to look at CDR3 sampling reproducibility and to determine if amplification provided any advantage in CDR3 reproducibility. For the KSU unamplified data set, two C57BL/6J mouse spleen RNA samples (#32 and #39) were sequenced independently two times each and the CDR3s sampled were compared. In the KSU data set, 32-1 shared 28% of its total unique reads with 32-2. (Figure 3). 32-2 shared 24% of its reads with 32-1 (Figure 3). KSU data set, 39, showed 25% overlap of their total unique reads between each sampling (Figure 3). For the Com2 data, since there was such a strong correlation between the sequence output between mRNA and TRNA samples of C57BL/6J spleen samples ($R^2 = 0.9644$), we considered these technical replicates. The mRNA data set shared 24% of its sequences with the TRNA data set while the TRNA data set shared 30% of its sequences with the mRNA data set. We also examined the resampling efficiency in the Com1 data set using the spleen mRNA-MMLV data sets that were reversed transcribed with two different primers, random hexamer and oligo-dT. Although this was not a perfect technical replicate, there was an $R^2$ of almost 0.94 in V-gene segments detected (Figure 1). Therefore, we felt these served as "incipient" technical replicates. The random hexamer data set shared 20% of the detected CDR3s and the oligo dT data set shared 32% of its CDR3 sequences. Therefore, regardless of data set, there was an average of 26-27% sampling overlap regardless of whether amplification was performed or not.

# Figure 3.
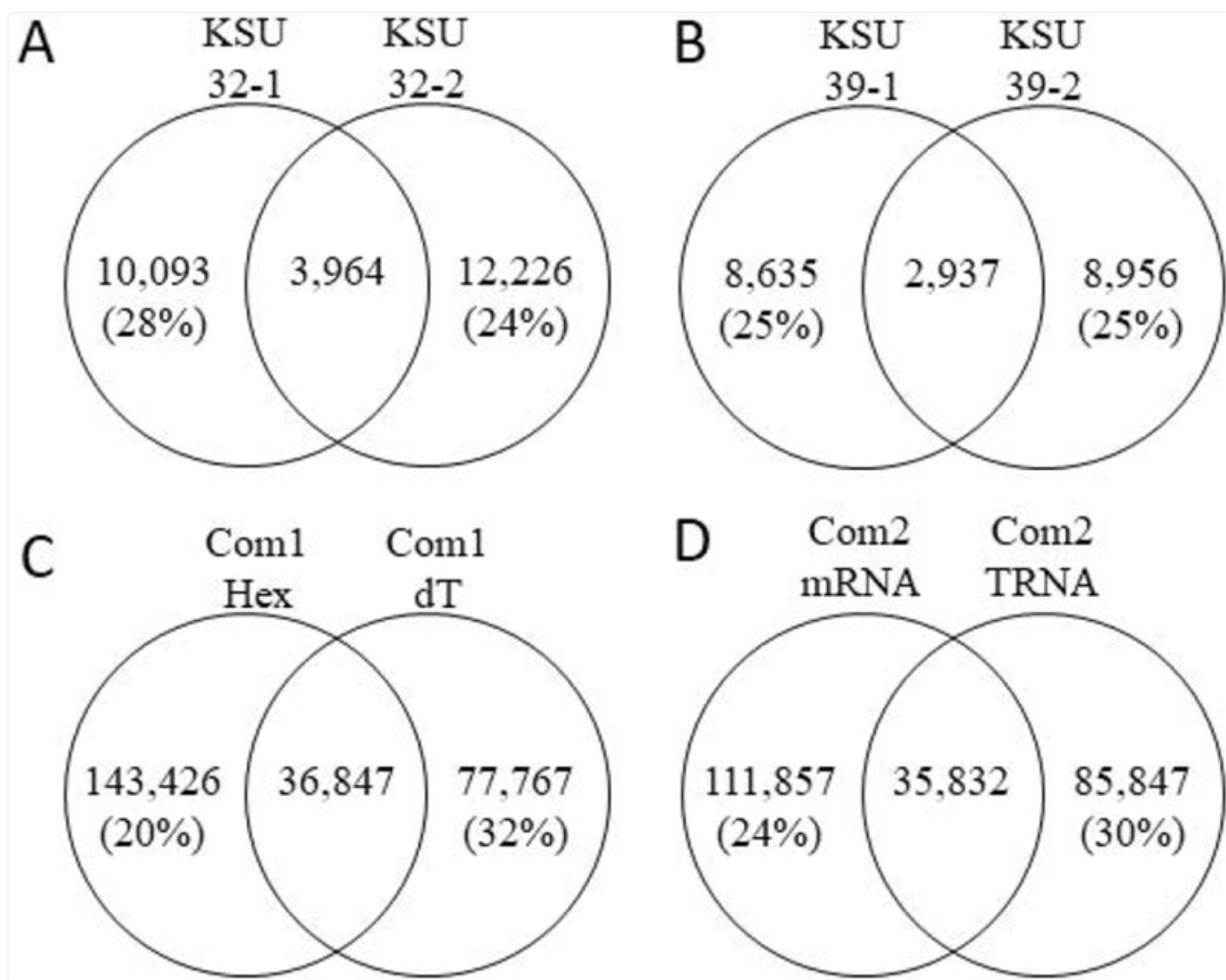
Overlap of CDR3 sequence detection between technical replicates. CDR3 amino acid sequences were compared between technical replicates. Sequences unique to one data set are displayed in the outer circles. Sequences shared between data sets are in the overlap. Percent of shared CDR3 sequences is displayed in parentheses in the outer circles. (A) KSU data sets 32-1 and 32-2. (B) KSU data sets 39-1 and 39-2. (C) Com1 data sets mRNA-MMLV-Hex and mRNA-MMLV-dT. (D) Com2 data sets mRNA and TRNA.

We also assessed the overlap in the detected CDR3s between the Com1 data set (mRNA-MMLV-Hex), Com2 data set (mRNA), and the KSU original data set to determine the extent of the overlap of CD3 detection using the different

methods. From the 295,116 CDR3 unique sequences that were detected, 2662 of those sequences were shared among all three data sets (Figure 4). The amplified data sets from Com1 and Com2 shared the most CDR3 sequences between them with 34,141 sequences found in both data sets (Figure 4). The KSU data set shared 59% of its detected CDR3 sequences with the Com1 and Com2 data sets (Figure 4). These data are consistent with the lower depth of sequencing of the unamplified data set compared to the Com1 and Com2 data sets where 19-32% overlap occurred in detected CDR3 sequences.

Figure 4.
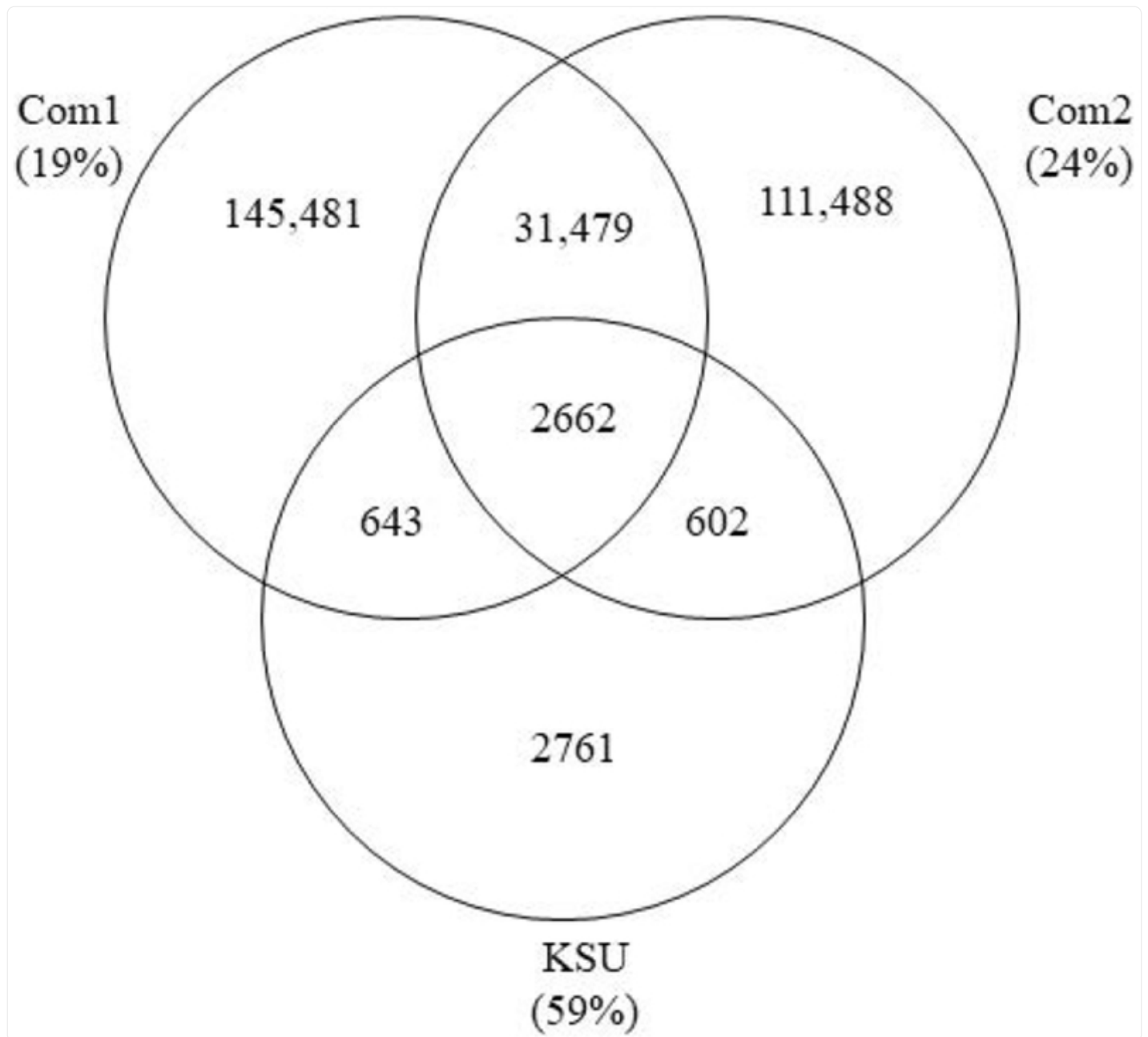


CDR3 sequence capture among Com1, Com2, and KSU data sets. CDR3 amino acid sequences were compared among the Com1 mRNA-MMLV-Hex, Com2 mRNA, and the KSU data sets. Percent of the repertoire shared with at least one other data set is listed in parentheses.

## Detection of High Frequency CDR3s

To gauge whether the highest frequency CDR3s can be detected by the different techniques, we assessed the 25 highest frequency CDR3s from each sequencing method. This resulted in a total of 48 unique CDR3s from the three different methods ([Figure 5](#)). The KSU data set detected all but one (CARGYFDVW) of these 48 sequences, the Com1 data set failed to detect four sequences (CARGTYW, CTWDEGNYW, CARGIYW, CARGSYW) and the Com2 data set detected all 48 sequences ([Figure 5](#)). The CDR3s that were not detected in the Com1 data set, did use V-gene segments that were detected in the data set. These data show that although the depth of sequencing of the KSU data set was about 10% of the amplified data sets, the data set still captured 98% of the highest frequency CDR3's.

# Figure 5.

| | KSU | Com1 | Com2 |
|---|---|---|---|
| CASVYDGYAFAYW | 4.2 | 0.1 | 1.3 |
| CARGAYW | 3.0 | 0.2 | 2.3 |
| CAREYDGYPYAMDYW | 1.5 | 0.1 | 1.3 |
| CARRWLHYAMDYW | 0.9 | 0.1 | 1.1 |
| CARGGYW | 0.8 | 0.0 | 0.6 |
| CARDYYGSSWYFDVW | 0.7 | 0.1 | 0.9 |
| CARTIYYGSSWFAYW | 0.7 | 0.1 | 0.2 |
| CAREELPHYFDYW | 0.5 | 0.0 | 0.0 |
| CARGGYDGYPHYAMDYW | 0.5 | 0.0 | 0.1 |
| CARGYYFDYW | 0.5 | 0.0 | 0.4 |
| CAKDYYGSSWYFDVW | 0.4 | 0.1 | 0.5 |
| CMRYSNYWYFDVW | 0.4 | 0.0 | 0.0 |
| CAREEGYYDWFAYW | 0.4 | 0.0 | 0.0 |
| CMRYSSYWYFDVW | 0.4 | 0.0 | 0.0 |
| CMRYGNYWYFDVW | 0.3 | 0.0 | 0.0 |
| CARRAHYYGSSYYFDYW | 0.3 | 0.0 | 0.1 |
| CASYSNYDYW | 0.3 | 0.0 | 0.3 |
| CARGTYW | 0.3 | ■ | 0.4 |
| CARWRVYYGNWYFDVW | 0.3 | 0.0 | 0.0 |
| CARDNWDWYFDVW | 0.3 | 0.0 | 0.1 |
| CARLYYYGSHWYFDVW | 0.3 | 0.0 | 0.1 |
| CARRDSNYLDYW | 0.3 | 0.0 | 0.2 |
| CASNAYYSNYVTPHFDYW | 0.2 | 0.0 | 0.3 |
| CARWYYGSSYEGYFDVW | 0.2 | 0.0 | 0.6 |

| | KSU | Com1 | Com2 |
|---|---|---|---|
| CTWDEGNYW | 0.2 | ■ | 0.3 |
| CASWDFAYW | 0.2 | 0.1 | 0.2 |
| CARSGGNWYFDVW | 0.1 | 0.0 | 0.2 |
| CASWEFAYW | 0.2 | 0.0 | 0.1 |
| CTRFAYW | 0.2 | 0.0 | 0.1 |
| CARGDYW | 0.2 | 0.0 | 0.2 |
| CARDDPFAYW | 0.2 | 0.0 | 0.1 |
| CARGGYYDYDWYFDVW | 0.2 | 0.0 | 0.0 |
| CARGGFAYW | 0.1 | 0.0 | 0.0 |
| CARSHYSAWFAYW | 0.1 | 0.0 | 0.1 |
| CARGFAYW | 0.0 | 0.0 | 0.0 |
| CARGYFDVW | ■ | 0.0 | 0.0 |
| CATYGSPFAYW | 0.1 | 0.0 | 0.0 |
| CARGGYSDYDWYFDVW | 0.2 | 0.0 | 0.0 |
| CARDYGSSYDSYW | 0.1 | 0.0 | 0.1 |
| CARSQNYWFAYW | 0.1 | 0.0 | 0.1 |
| CAKAPIYYDYDGVFYYAMDYW | 0.1 | 0.0 | 0.4 |
| CARSPDGYYAMDYW | 0.2 | 0.0 | 0.3 |
| CARSGAYYRDYYAMDYW | 0.2 | 0.0 | 0.3 |
| CARDGSHAMDYW | 0.1 | 0.0 | 0.3 |
| CARGIYW | 0.2 | ■ | 0.2 |
| CARQLYAMDYW | 0.1 | 0.0 | 0.2 |
| CARGSYW | 0.2 | ■ | 0.2 |
| CARRLDYW | 0.0 | 0.0 | 0.2 |

High frequency CDR3s detected among the Com1, Com2, and KSU data sets. The top 25 CDR3s from each data set (48 total) were compiled and percent of repertoire compared. Black boxes represent no detected reads (true zero). Rounded zeros are represented as 0.0.

## CDR3 Frequency Assessment

While the high frequency CDR3s were shared among at least two sequencing runs, most of the CDR3 sequences detected were unique to each sequencing run (Figure 4). To determine the frequency of unique CDR3 sequences, we

compared the most frequent, least frequent, and average percent of the measured CD3 repertoire (Table 5). The highest frequency CDR3 was detected at 4.16%, 0.22%, 2.26%, and of the repertoire for the KSU, Com1, and Com2 data sets, respectively (Table 5). The lowest frequency CDR3s, representing only a single detected read, were 0.0088%, 0.0004%, and 0.0002%, of the repertoire for the KSU, Com1, and Com2 data sets, respectively (Table 5). The average detection level for the KSU data set was 0.015%, 0.0006% for the Com1 data set, and 0.0007% for the Com 2data set (Table 5).

## Table 5:

| | Whole Repertoire | | | Unique Repertoire | | |
|---|---|---|---|---|---|---|
| | KSU | Com1 | Com2 | KSU | Com1 | Com2 |
| Minimum | 0.008758 | 0.000446 | 0.000221 | 0.008758 | 0.000446 | 0.000221 |
| Maximum | 4.1655 | 0.216969 | 2.259680 | 0.035032 | 0.005804 | 0.026044 |
| Average | 0.014997 | 0.000555 | 0.00684 | 0.008929 | 0.000478 | 0.00290 |

Open in a new tab

We also examined the frequency of CDR3s that were unique to each data set. Overall, the maximum and the average frequencies of the data sets were reduced compared to the whole repertoire (Table 5). This demonstrates that the unique reads in each data set were most likely transcripts from low frequency B cells. Moreover, these data suggest that even without amplification, the KSU data set was detecting the most prevalent CDR3s and many low-frequency sequences.

# Discussion

Illumina sequencing of total RNA from mouse spleen is able to capture a representative sample of the splenic B cell repertoire. This snapshot of the repertoire, while producing less reads than amplified data sets, detected more V-gene segments than data sets that used two different amplification strategies and captures 98% of the high frequency CDR3s found in the amplified data sets. While amplified data sets provide more CDR3 depth of coverage, the unamplified data sets produced from an RNASeq allow for further data mining, eliminate as much primer bias as possible and provides an accurate representation of the repertoire.

Sequencing requirements of the B cell receptor are more challenging than those of the T cell receptor. There are no consensus sequences to reference (10). Additionally, transcripts from the germline which are not successfully rearranged

can be detected (23). Therefore, avoiding bias is one of the main priorities for antibody repertoire sequencing (16). PCR errors are accumulated through the amplification process which can falsely inflate the repertoire or they can add suspected mutations that do not exist (10, 17, 24, 25) and they may not be distinguishable from low level mutations that actually do exist in the repertoire (24, 26). PCR biases can be introduced because of primer binding properties, CG content, mispriming, non-specific binding, and errors during replication (16, 27–29). A specific issue in targeting antibody gene segments is primer annealing efficiencies since the gene segments that make up the murine IgH locus are similar, though not identical (28). The biases inherent in the multiplex PCR can lead to false repertoire skewing, gene frequency inaccuracies, and a less comprehensive view of the repertoire (25, 30, 31). The development of these multiplex primers, is highly challenging (32). Work by Bashford-Rogers (33) shows that there is little difference between RNA-capture, 5'RACE or PCR amplification in humans but others (27) have shown that PCR does create biases and 5' RACE helps reduce these.

Our results demonstrate some of the issues of assessing B-cell repertoires using massively multiplexed PCR reactions. While reproducibility for technical replicates was moderate to high, there was a large range of total read numbers across methodologies. Given the increase in read results with mRNA samples compared to TRNA samples, the authors hypothesize that mRNA increases the availability of antibody specific transcripts, increasing the overall read counts. Additionally, the reverse transcriptase plays a role in sequence generation. AMV reverse transcriptase has an RNase H activity, where MMLV reverse transcriptase does not (34). This activity may cleave RNA transcripts prior to completion of amplification, resulting in shorter reads with inadequate length for amplification. The use of oligo-dT primers will begin reverse transcription from the most 3' end of the transcript. The target of amplification is around 1.5 kilobases upstream of the poly-A tail (35), likely leading to shorter reads and failure to amplify the region of interest. Random hexamer primers maybe able to overcome this bias due to their random priming nature.

In addition to variations in read depth, the correlations of these data sets to the unamplified KSU data sets were low to moderate, even when the same bioinformatics processing was employed. Of significant concern is the Com1 and Com2 sets failed to detect 13 V-gene segments that were detected by the unamplified KSU RNASeq data set. The absence of these genes draws into question how one compares the various data sets with correlation coefficients that are below 0.7 when technical replicates of the same sequencing are greater than 0.99. Although Carlson *et al*. argues that amplification methodologies can capture the entire repertoire, there are concerns (11). Even when we only looked at the V-gene families detected at the highest frequency there were omissions. Of the 34 V-genes that we categorized as "high frequency" (>1% of the repertoire), Com1 found lower detection levels (defined as less than two-fold that found in the KSU data set) for five gene segments (V1-26, V1-18, V1-50, V4-1, and V2-6) and nine (V9-3, V4-1, V2-6, V1-, V5-17, V2-2, V8-8, V11-2, V14-2) for the Com2 data set. These results suggest that those methods are skewing the reported repertoire by missing or underreporting those high frequency V-gene segments. Interestingly, two V-gene segments (V4-1 and V2-6) were underrepresented in both data sets, but the other V-gene segments were unique.

The failure of primers to capture specific V-genes is not a new discovery, as primers failed to adequately sequence

hybridomas previously ([18](#), [36](#)). "Universal" primers for the human antibody repertoire do exist, but some questions remain if they cover the entire repertoire ([37](#)). The difficulty in developing a "universal" or even highly comprehensive primer set for the mouse is likely due to their highly varied leader sequences, V-gene segments, and framework regions. Primer design would have to rely on massively multiplexed reactions and/or degenerate primers. Additionally, the most commercially viable amplification methods would need to amplify across multiple common strains adding additional levels of complexity. Indeed, in our attempts to design "universal" primers, we found a minimum of 11 primer sets would be needed to detect V-genes associated for each isotype. Even then, there were still issues with matching PCR conditions and efficiency. Methods to overcome the biases detected in amplification have been developed, such as the use of 5' RACE ([26](#), [38](#), [39](#)) and using molecular barcodes or identifiers ([40](#), [41](#)). However, these methods are expensive and have their own draw backs. Replication of the entire repertoire using 5'RACE would still require the use of multiple constant region primers, leading to the same multiplexing issues. Barcoding can have errors and chimeric reads making repertoires difficult to reconstruct ([42](#)). This latter issue is not a problem with our RNASeq data.

While bias exists in the massively multiplexed amplification process, there may be some sequencing errors in the unamplified KSU data set as well. While not specifically amplified for antibody sequences, random hexamers and oligo-dT capture beads are used prior to sequencing to generate the library ([43](#)) and some biases have been observed in random hexamer binding ([44](#)). The use of oligo-dTs can result in enhancement of the 3' end of transcripts ([45](#)). We do not think this is particularly problematic since Illumina sequence methodology aims to reduce bias in their library preparations by combining the random hexamer and oligo-dT capture. Additionally, while all libraries were sequenced on the Illumina platform, 3.19% of high-quality Illumina reads contain false base calls, which are impossible to differentiate using normal bioinformatics methods ([10](#)). Over representation of specific dineuclotides can also be detected in sequencing which are not related to primer usage ([46](#)). Therefore, although we hope to reduce bias and omission by doing RNASeq, we still have some technical issues that keep the data set from being a perfect reflection of the repertoire. Multiple technical replications help reduce the impact of this problem.

One analysis that was not pursued in this current investigation was the identification of clonally related sequences. We were specifically interested in the functional antibody repertoire present in the spleen and focused our analysis at the transcript level. We acknowledge that the overrepresentation of some sequences may be likely within our data sets since we don't use barcoding or clonality analysis to collapse similar mRNA sequences. Nevertheless, the overrepresentation of a specific sequence by an overly productive cell is also representative of cellular activation of transcription, and likely, functional antibody protein in the body ([47](#), [48](#)).

The lack of amplification also results in varying sequence lengths in our data set. We selected 40nt as our minimum cutoff to provide us enough information to detect V-, D-, and J- gene segments. While some short sequences were included in the data analysis, our overall average sequence length was 287nt, with productive sequences averaging 331nt and unknown sequences averaging 270 nt (data not shown). Overall, less than 0.5% of the sequences analyzed where less than 100nt long (data not shown). Therefore, we do not think that sequence length is an issue in this study.

Although there are issues using massively multiplexed PCR reactions, there are advantages that may overshadow the disadvantages. For example, increased sequencing depth (*e.g.* 1,260,000 reads *vs.* 11,200 complete reads), better low frequency CDR3 detection (20-fold more unique CDR3 sequence than the unamplified KSU data set), and sequencing costs can be lower than an Illumina MiSeq run (by integrating multiplexing/barcoding).

Although amplification provides more detail in the CDR3 repertoire, if one is interested in the B cell clones that are most prevalent, then RNASeq does not appear to be at a disadvantage. CDR3 resampling was similar (20-32%) regardless of method. Additionally, when examining high frequency CDR3s, the unamplified data set only failed to detect a single unique CDR3 sequence, while the Com1 data set failed to detect four. Without using an amplification method with unique barcoding, such as that used in Shugay et al ([49](#)) it is impossible to tell which unique CDR3s are, in fact, correctly identified new sequences and which involve miscalls, leading to false diversity. As our current paper does not focus on the actual diversity of the repertoire, and instead focuses on the differences among sample preparations and sequencing methods, we do not attempt to identify false call unique CDR3s as doing so may falsely reduce diversity. Instead, the we used strict read quality cutoffs to prevent low quality base calls from being included in analysis, and we point to our highly overlapping V-gene segment usage ($R^2 = 0.9995$ among multiple sequencing replicates to assure us that we minimized artifacts. Additionally, all sequencing was performed on the Illumina platform, so false base call rates should be similar across methodologies which we were focused on in this study.

When preparing for antibody repertoire sequencing, multiple factors must be considered within the framework of the specific biological questions being asked. This includes needed coverage, cost, and starting material ([39](#)). Additionally, it is important to consider that all repertoire sequencing is merely a snapshot of a constantly shifting image ([25](#)). We will also never be able to fully capture the full diversity of the B cell immunoglobulin repertoire, which is estimated to range from $10^6$-$10^7$ possible unique rearrangements and mutations ([25](#), [39](#)) to as much as $10^{13}$ ([15](#)). The failure of the KSU data set to detect rare clones compared to the amplified data sets is likely due to this; but even the amplified data sets only sampled a fraction of the total repertoire. Therefore, one must decide how "deep" is adequate for the question being addressed.

During this investigation, we also had to address the issue that starting material may influence the quality of one's sequencing. mRNA as a starting template increased reads up to two orders of magnitude. Furthermore, random hexamer primers and MMLV reverse transcriptase generally yielded higher read count results. Interestingly, the use of mRNA, with MMLV reverse transcriptase and hexanucleotide primers is most technically like that used in Illumina sequencing. However, additional data will be needed to confirm our observations since we did not pursue this aspect of the study in detail and the replicate number did not allow for robust statistical comparisons.

In conclusion, we have demonstrated that sequencing of unamplified splenic RNA provides a realistic snapshot of the total splenic B cell repertoire. We also have demonstrated that a good understanding of the bioinformatics work flow and reporting of the methodology is critical and cannot be understated. We understand that there are cellular biases and

transcript stability differences with in B cell subpopulations ([50](#), [51](#)). However, for the purpose of assessing a whole tissue B cell repertoire, unamplified RNASeq can provide a glimpse of the most prevalent B cell clones. The unamplified approach could just as well be applied to specific cell populations when the application requires it. Moreover, an unamplified data set may detect V-gene segments that amplified data sets miss.

## Supplementary Material

Supp FigS1

[NIHMS987236-supplement-Supp_FigS1.pdf](#) (365.2KB, pdf)

## Acknowledgements

## Abbreviations:

**V**

Variable Gene Segment

**D**

Diversity Gene Segment

**J**

Joining Gene Segment

**TRNA**

Total RNA

**AMV**

Avian Myeloblastosis Virus

**MMLV**

Moloney Murine Leukemia Virus

**IMGT**

ImMunoGeneTics

# References

1. Ademokun A, Wu Y-C, Martin V, Mitra R, Sack U, Baxendale H, Kipling D, and Dunn Walters D (2011) Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. Aging cell 10, 922–930 [DOI] [PMC free article] [PubMed] [Google Scholar]

2. Galson JD, Clutterbuck EA, Truck J, Ramasamy MN, Munz M, Fowler A, Cerundolo V, Pollard AJ, Lunter G, and Kelly DF (2015) BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. Immunology and cell biology 93, 885–895 [DOI] [PMC free article] [PubMed] [Google Scholar]

3. Strauli NB, and Hernandez RD (2016) Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome medicine 8, 60. [DOI] [PMC free article] [PubMed] [Google Scholar]

4. Fan Q, Nelson CS, Bialas KM, Chiuppesi F, Amos J, Gurley TC, Marshall DJ, Eudailey J, Heimsath H, Himes J, Deshpande A, Walter MR, Wussow F, Diamond DJ, Barry PA, Moody MA, Kaur A, and Permar SR (2017) Plasmablast Response to Primary Rhesus Cytomegalovirus (CMV) Infection in a Monkey Model of Congenital CMV Transmission. Clinical and vaccine immunology: CVI 24, e00510–00516 [DOI] [PMC free article] [PubMed] [Google Scholar]

5. Nivarthi UK, Kose N, Sapparapu G, Widman D, Gallichotte E, Pfaff JM, Doranz BJ, Weiskopf D, Sette A, Durbin AP, Whitehead SS, Baric R, Crowe JE Jr., and de Silva AM (2017) Mapping the Human Memory B Cell and Serum Neutralizing Antibody Responses to Dengue Virus Serotype 4 Infection and Vaccination. Journal of virology 91, e02041–02016 [DOI] [PMC free article] [PubMed] [Google Scholar]

6. Racanelli V, Brunetti C, De Re V, Caggiari L, De Zorzi M, Leone P, Perosa F, Vacca A, and Dammacco F (2011) Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. PloS one 6, e25606. [DOI] [PMC free article] [PubMed] [Google Scholar]

7. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, Pui CH, and Campana D (2012) Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. Blood 120, 5173–5180 [DOI] [PMC free article] [PubMed] [Google Scholar]

8. Gerasimov E, Zelikovsky A, Mandoiu I, and Ionov Y (2017) Identification of cancer-specific motifs in mimotope profiles of serum antibody repertoire. BMC bioinformatics 18, 33–38 [DOI] [PMC free article] [PubMed] [Google Scholar]

9. Rettig TA, Ward C, Pecaut MJ, and Chapes SK (2017) Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station. Gravitational and Space Research 5, 2–23 [PMC free article] [PubMed] [Google Scholar]

10. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, Turchaninova MA, Lukyanov S, Lebedev YB, and Chudakov DM (2012) Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. European journal of immunology 42, 3073–3083 [DOI    ] [PubMed] [Google Scholar    ]

11. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, Wu D, Wood BL, Rieder MJ, and Robins H (2013) Using synthetic templates to design an unbiased multiplex PCR assay. Nature communications 4, 2680. [DOI    ] [PubMed] [Google Scholar    ]

12. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, and Koralov S (2011) High-resolution description of antibody heavy-chain repertoires in humans. PloS one 6, e22365. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

13. Bashford-Rogers RJM, Palser A, Huntly B, Rance R, Vassiliou G, Follows G, and Kellam P (2013) Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. Genome research 23, 1874–1884 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

14. Briney B, Le K, Zhu J, and Burton DR (2016) Clonify: unseeded antibody lineage assignment from next-generation sequencing data. Scientific reports 6, 23901. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

15. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, Valai A, Lopes T, Radbruch A, Winkler TH, and Reddy ST (2017) Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell reports 19, 1467–1478 [DOI    ] [PubMed] [Google Scholar    ]

16. Best K, Oakes T, Heather JM, Shawe-Taylor J, and Chain B (2015) Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. Scientific reports 5, 14629. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

17. van Dijk EL, Jaszczyszyn Y, and Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. Experimental cell research 322, 12–20 [DOI    ] [PubMed] [Google Scholar    ]

18. Wang Y, Chen W, Li X, and Cheng B (2006) Degenerated primer design to amplify the heavy chain variable region from immunoglobulin cDNA. BMC bioinformatics 7 Suppl 4, S9. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

19. Alamyar E, Duroux P, Lefranc M-P, and Giudicelli V (2012) IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations:

IMGT/V-QUEST and IMGT/HighV-QUEST for NGS In Immunogenetics (Christiansen FT, and Tait BD, eds) Vol. 882 pp. 569–604, Humana Press; [DOI    ] [PubMed] [Google Scholar    ]

20. Rettig TA, Ward C, Bye BA, Pecaut MJ, and Chapes SK (2018) Characterization of the naive murine antibody repertoire using unamplified high-throughput sequencing. PloS one 13, e0190982. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

21. (2017) IMGT Repertoire (IG and TR).

22. (2017) Igh immunoglobulin heavy chain complex [Mus musculus (house mouse)] - Gene - NCBI.

23. Angelin-Duclos C, and Calame K (1998) Evidence that immunoglobulin VH-DJ recombination does not require germ line transcription of the recombining variable gene segment. Molecular and cellular biology 18, 6253–6264 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

24. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, Kirgizova VI, Merzlyak EM, Staroverov DB, Bolotin DA, Mamedov IZ, Izraelson M, Logacheva MD, Kladova O, Plevova K, Pospisilova S, and Chudakov DM (2016) High-quality full-length immunoglobulin profiling with unique molecular barcoding. Nature protocols 11, 1599–1616 [DOI    ] [PubMed] [Google Scholar    ]

25. Hou XL, Wang L, Ding YL, Xie Q, and Diao HY (2016) Current status and recent advances of next generation sequencing techniques in immunological repertoire. Genes and immunity 17, 153–164 [DOI    ] [PubMed] [Google Scholar    ]

26. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, and Quake SR (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nature biotechnology 32, 158–168 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

27. Benichou J, Ben Hamo R, Louzoun Y, and Efroni S (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135, 183–191 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

28. Boyd SD, and Joshi SA (2014) High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiology spectrum 2, AID-0017–2014 [DOI    ] [PubMed] [Google Scholar    ]

29. Robasky K, Lewis NE, and Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. Nature reviews. Genetics 15, 56–62 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

30. Langerak AW, Bruggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D, Cazzaniga G, Giudicelli V, Lefranc MP, Giraud M, Macintyre EA, Hummel M, Pott C, Groenen P, and Stamatopoulos K (2017) High-Throughput Immunogenetics for Clinical and Research Applications in Immunohematology:

Potential and Challenges. Journal of immunology (Baltimore, Md. : 1950) 198, 3765–3774 [DOI ] [PubMed] [Google Scholar ]

31. Baum P, Venturi V, and Price D (2012) Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. European journal of immunology 42, 2834–2839 [DOI ] [PubMed] [Google Scholar ]

32. Calis JJ, and Rosenberg BR (2014) Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends in immunology 35, 581–590 [DOI ] [PMC free article] [PubMed] [Google Scholar ]

33. Bashford-Rogers RJ, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, Douek DC, Vassiliou GS, Follows GA, Hubank M, and Kellam P (2014) Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC immunology 15, 29. [DOI ] [PMC free article] [PubMed] [Google Scholar ]

34. Kotewicz ML, Sampson CM, D'Alessio JM, and Gerard GF (1988) Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. Nucleic acids research 16, 265–277 [DOI ] [PMC free article] [PubMed] [Google Scholar ]

35. (2018) Ighm immunoglobulin heavy constant mu [Mus musculus (house mouse)] - Gene - NCBI. Pubs

36. Berdoz J, Monath TP, and Kraehenbuhl JP (1995) Specific amplification by PCR of rearranged genomic variable regions of immunoglobulin genes from mouse hybridoma cells. PCR methods and applications 4, 256–264 [DOI ] [PubMed] [Google Scholar ]

37. Sun Y, Liu HY, Mu L, and Luo EJ (2012) Degenerate primer design to clone the human repertoire of immunoglobulin heavy chain variable regions. World journal of microbiology & biotechnology 28, 381–386 [DOI ] [PMC free article] [PubMed] [Google Scholar ]

38. Lees WD, and Shepherd AJ (2017) Studying Antibody Repertoires with Next-Generation Sequencing. Methods in molecular biology (Clifton, N.J.) 1526, 257–270 [DOI ] [PubMed] [Google Scholar ]

39. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, Mora T, Thomas-Vaslin V, Walczak AM, and Boudinot P (2013) The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. Frontiers in immunology 4, 413. [DOI ] [PMC free article] [PubMed] [Google Scholar ]

40. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, and Vollmers C (2016) Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier-Guided Amplicon Assembly. Journal of immunology (Baltimore, Md. : 1950) 196, 2902–2907 [DOI ]

[PubMed] [Google Scholar    ]

41. Robinson WH (2015) Sequencing the functional antibody repertoire--diagnostic and therapeutic discovery. Nature reviews. Rheumatology 11, 171–182 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

42. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, and Pevzner PA (2017) Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Reads. Journal of immunology (Baltimore, Md. : 1950) [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

43. Hansen KD, Brenner SE, and Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research 38, e131. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

44. van Gurp TP, McIntyre LM, and Verhoeven KJ (2013) Consistent errors in first strand cDNA due to random hexamer mispriming. PloS one 8, e85583. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

45. Weber CF, and Kuske CR (2012) Comparative assessment of fungal cellobiohydrolase I richness and composition in cDNA generated using oligo(dT) primers or random hexamers. Journal of microbiological methods 88, 224–228 [DOI    ] [PubMed] [Google Scholar    ]

46. Zheng W, Chung LM, and Zhao H (2011) Bias detection and correction in RNA-Sequencing data. BMC bioinformatics 12, 290. [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

47. Lou Z, Casali P, and Xu Z (2015) Regulation of B Cell Differentiation by Intracellular Membrane-Associated Proteins and microRNAs: Role in the Antibody Response. Frontiers in immunology 6 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

48. Zhang Q, Bernatoniene J, Bagrade L, Paton JC, Mitchell TJ, Hammerschmidt S, Nunez DA, and Finn A (2006) Regulation of Production of Mucosal Antibody to Pneumococcal Protein Antigens by T-Cell-Derived Gamma Interferon and Interleukin-10 in Children. Infection and Immunity 74, 4735–4743 [DOI    ] [PMC free article] [PubMed] [Google Scholar    ]

49. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva EV, Plevova K, Linnemann C, Shagin D, Pospisilova S, Lukyanov S, Schumacher TN, and Chudakov DM (2014) Towards error-free profiling of immune repertoires. Nature methods 11, 653–655 [DOI    ] [PubMed] [Google Scholar    ]

50. Allman D, Lindsley RC, DeMuth W, Rudd K, Shinton SA, and Hardy RR (2001) Resolution of three nonproliferative immature splenic B cell subsets reveals multiple selection points during peripheral B cell maturation. Journal of immunology (Baltimore, Md. : 1950) 167, 6834–6840 [DOI    ] [PubMed] [Google Scholar    ]

51. Allman DM, Ferguson SE, Lentz VM, and Cancro MP (1993) Peripheral B cell maturation. II. Heat-stable antigen(hi) splenic B cells are an immature developmental intermediate in the production of long-lived marrow-derived B cells. Journal of immunology (Baltimore, Md. : 1950) 151, 4431–4444 [PubMed] [Google Scholar    ]

## Associated Data

*This section collects any data citations, data availability statements, or supplementary materials included in this article.*

## Supplementary Materials

Supp FigS1

NIHMS987236-supplement-Supp_FigS1.pdf (365.2KB, pdf)